The interexaminer reproducibility of physical examination of the cervical spine

Pool, J.J.M.; Hoving, J.L.; de Vet, H.C.W.; van Mameren, H.; Bouter, L.M.

published in
Journal of Manipulative and Physiological Therapeutics
2004

DOI (link to publisher)
10.1016/j.jmpt.2003.12.002

document version
Publisher's PDF, also known as Version of record

The Interexaminer Reproducibility of Physical Examination of the Cervical Spine

Jan J. Pool,[a] Jan L. Hoving, PhD,[a] Henrica C. de Vet, PhD,[a] Henk van Mameren, MD, PhD,[b] and Lex M. Bouter, PhD[a]

Abstract

Objective: To assess the interexaminer reproducibility of physical examination of the cervical spine.

Methods: Two physiotherapists independently judged the general mobility and the intersegmental mobility (segments C0-T2) of the neck and the pain that was provoked. Percentage agreement and Cohen's κ expressed agreement of dichotomous variables; limits of agreement expressed agreement of continuous variables; and intraclass correlation coefficients (ICCs) expressed the reliability of continuous variables.

Results: Agreement for general mobility showed κ between 0.05 and 0.61, and for the intersegmental mobility, it showed κ values between −0.09 and 0.63. Agreement for provoked neck pain within 1 point of an 11-point numerical rating scale (NRS) varied between 46.9% and 65.7% for general mobility and between 40.7% and 75.0% for intersegmental mobility. The ICCs varied between 0.36 and 0.71 for general mobility and between 0.22 and 0.80 for intersegmental mobility.

Conclusions: Despite the use of a standardized protocol to assess general mobility and intersegmental mobility of the cervical spine, it is difficult to achieve reasonable agreement and reliability between 2 examiners. Likewise, the patients are not able to score the same level of provoked pain in 2 assessments with an interval of 15 minutes. (J Manipulative Physiol Ther 2004;27:84-90)

Key Indexing Terms: *Agreement; Cervical Spine; Mobility; Reliability; Reproducibility*

Introduction

Neck pain is a common complaint in the general population, and its point prevalence is around 15%.[1] Patients with neck pain who consult their general practitioner usually receive advice and analgesics, and approximately 43% are referred to a physical therapist or manual therapist.[1] Manual assessment of the mobility of the cervical spine is made by many professionals, including physical therapists, manual therapists, chiropractors, and physicians. The physical examination of the cervical spine is based on the assessment of passive and/or active range of movement, including possible pain provocation during or at the end of the range of movement.[2] This is assumed to provide important information with regard to the patient's impairments. Moreover, the results of the examination and the patient's pain response during the examination are the basis for the proposed treatment, and the results of the physical examination will be used to evaluate the treatment results.[2,3] Therefore, it is important to know the reproducibility of these assessments. The reproducibility can be studied in terms of reliability and agreement. *Reliability* is defined as the ability to differentiate between patients and *agreement* is defined as the extent to which observers obtain the same measurement values in a test.[4-7]

Many techniques are used by physical or manual therapists to examine the cervical spine,[3,8] but the reproducibility of these techniques is questionable.[9-13] Several studies have drawn different conclusions with regard to the reproducibility of manual assessment techniques.[9-13] The majority of these studies report that better operational definitions and testing procedures are needed.

The aim of this study is to investigate the interobserver reliability and agreement of a standardized physical examination for patients with nonspecific neck pain based on a protocol in which manual techniques are used to assess the general and intersegmental mobility of the cervical spine. A new approach in this study, in comparison with former

[a]Institute for Research in Extramural Medicine, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands, and Medical Centre Impact, Zoetermeer, The Netherlands.

[b]Department of Anatomy/Embryology, Faculty of Medicine, Maastricht University, Maastricht, The Netherlands.

Submit requests for reprints to: Jan J. Pool, PT, BSc, Institute for Research in Extramural Medicine, Vrije Universiteit Medical Centre, Van der Boechorststraat 7, 1081 BT Amsterdam, The Netherlands (e-mail: *j.pool@vumc.nl*).

Paper submitted December 2, 2002.

0161-4754/$30.00

doi:10.1016/j.jmpt.2003.12.002

studies, is to include assessment of the interobserver reproducibility of the patient's pain response to the testing procedures, reported on an 11-point numerical rating scale (NRS).

Methods

During a period of 4 months (April 1999 to June 1999), 32 patients were invited to participate in the study. Patients were referred by local general practitioners in the city of Zoetermeer, in the Netherlands, to a practice providing physical and manual therapy. The patients had a similar profile to those who had participated in a recently completed randomized clinical trial on patients with neck pain.[14]

Two experienced physical therapists (JJMP and LA) performed the examination of the cervical spine. The physical therapists were trained in the use of the protocol, and the movements included in the tests are part of the routine therapy they provide for patients with neck pain. They also followed the standardized protocol, which describes the performance of each of the tests in detail.

Measurements

Data on demographics; patient characteristics (duration, previous episodes, number of episodes); pain on a visual analogue scale; and disability, using the Neck Pain Disability index,[15] were collected prior to the assessment.

General Mobility of the Cervical Spine

The standardized clinical assessment of the general mobility of the cervical spine consists of 6 movements:

1. Full flexion and extension;
2. High cervical flexion (nodding) and extension (C0-1);
3. Rotation to the right and to the left;
4. Lateral flexion to the right and to the left;
5. Combined movement A: rotation to the right and to the left, combined with extension and homolateral flexion (combination of the entire available movement in rotation, lateral flexion, and extension); and
6. Combined movement B: lateral flexion to the right and to the left, combined with heterolateral rotation (isolating high cervical rotation).

At the end of the voluntary movement, the examiner applied a gentle passive pressure to guide the patient's movement to the end of range to obtain a clear estimation of the range, the tissue resistance to movement, and any pain response.[8] The examiner classified the movements as limited or not limited, and the patient was asked to score the provoked pain at the end of each movement on an 11-point numerical rating scale, ranging from 0 (no pain at all) to 10 (extremely painful). For all movements, the patient was seated on a chair with the hands on the thighs and the back against the backrest.

Intersegmental Mobility of the Cervical Spine

The passive segmental assessment, from segment C0 to segment T2, was made with the patient in a supine position and the examiner sitting behind the patient. The commonly used technique to assess the segments C2 to T2 included fixation of the lower segmental level and lateral flexion to the right and to the left. Rotation was used to assess the level C1-2 and flexion was used for the segment C0-1, because these segments have different movement potentials.[3] The examiner classified the movement as limited or not limited and the patient was asked to score the provoked pain at the end of each movement on an 11-point numerical rating scale.

The examiners were trained in the assessment protocol, and the order of the examination was randomized according to a computer-generated random sequence table. The time interval between the assessments was approximately 15 minutes. The examiners were blinded to each other's results and had no contact with each other between the assessments. In the absence of the other examiner's results, a research assistant registered the assessment.

Assessment of Reproducibility

Reproducibility was quantified in 2 ways: by measures of agreement, such as κ and the Bland-Altman method,[4,5,7,16] and measures of reliability, such as the intraclass correlation coefficient (ICC).[6,7,17] Agreement measures assess the absolute agreement and try to quantify the measurement error. Reliability parameters assess how well persons can be distinguished from each other despite measurement errors.[7] As the mobility scores were dichotomous (limited or not limited), Cohen's κ was used to calculate the agreement.[16] A κ score of 0.40 or higher was considered to be acceptable.[6,7,17] The provoked pain scores on the numerical rating scale were analyzed as continuous variables. The Bland-Altman method was used to assess the agreement, ie, the extent to which examiners obtained the same measurement values in a test.[4] Using the Bland-Altman method, the interobserver difference was calculated and plotted against the mean of the 2 measurements. The magnitude of the difference between the mobility scores and their distribution were visualized. The standard deviation of the difference gives an indication of the agreement of the 2 measurements. The 95% limits of agreement were calculated (difference $\pm$ $1.96 \times SD_{difference}$), which gives an indication of the total error, ie, bias and random error.[18] The percentage of agreement of the measurements was determined, allowing both 1 and 2 points difference on the numerical rating scale (0-10). For the continuous variables, the intraclass correlation coefficient was calculated as a measure of reliability, representing the ability to distinguish between patients.[6,7,17,19] The ICC is based on a 2-way random effect model ICC (2.1),[6] with the observers as a random factor (see Appendix), focusing on agreement, and ranges between 0 (no reliability) to 1 (perfect reliability). A cutoff point of ICC >

Table 1. *Characteristics of patients*

Characteristics		SD
Female (%)	62.5	
Previous episodes of neck complaints (%)	56.3	
Most important complaints (%)*		
Pain	78.1	
Limitation of movement	40.6	
Stiffness	28.1	
Mean age (years)	45.5	9.2
Mean number of episodes in the previous 5 years	8.6	22.1
Pain score†		
Mean	5.2	2.0
Maximum	7.2	2.4
Present	4.2	2.3
Duration of neck pain (weeks)	13.5	
Mean NDI score‡	15.2	8.3

N = 32.
SD, standard deviation; *NDI*, Neck Disability Index.
*Maximum 3 complaints.
†Pain was measured on a numerical 11-point scale ranging from 0 = no pain to 10 = worst pain.
‡Neck Disability Index; disability and pain measured by 10 items ranging from 0 to 5 points; maximum disability 50 points.

0.75 was chosen a priori as an indication of acceptable reliability.[20]

Results

Patient Characteristics

During a period of 4 months (April 1999 to June 1999), 32 consecutive patients with nonspecific neck pain were included. The mean age of the patients was 45 years, and approximately 63% were female patients (Table 1). The patients had suffered from neck pain for a median duration of 13 weeks and the neck pain was recurrent in over 50% of patients. Pain was the most important complaint in 78% of the patients and limitation of movement was the most important complaint in 40.6 % of the patients; the patients were allowed to score a maximum of 3 complaints. The patients rated the severity of their current neck pain, on average, as 4.2 points on an 11-point numerical rating scale. The maximum pain score in the last week was, on average, 7.2 points. The mean score for the Neck Disability Index (NDI) was 15.2 points.

Interexaminer Reproducibility of Cervical Mobility

The data on the general mobility are presented in Table 2. The observers scored each movement as limited or not limited. The prevalence of limited movements during the examination of the cervical spine varied from 1.6 % for the high flexion end extension movements to 64.5% for the combined movement to the right. The agreement for the general cervical movement ranged from 52% to 97%, with a mean of 71%. The κ for general cervical movement ranged from 0.05 to 0.61. Only rotation to the left and the combined movement A to the left showed a κ value higher than 0.40.

Table 3 presents the data on intersegmental mobility. The prevalence of limited segmental movements varied from 8.1% for the intersegmental movement on the level C1-2 on the right to 56.5% for the intersegmental movement on the level C4-5 on the right. The agreement for the intersegmental movements varied from 48% to 90%, with a mean of

Table 2. *Interexaminer agreement of general mobility of the cervical spine*

	Limited movements* (mean prevalence) %	Agreement %	Kappa
Flexion	21.0	71	0.19
Extension	33.9	71	0.39
High flexion C0-1	1.6	97	-†
High extension C0-1	1.6	97	-†
Rotation right	45.2	61	0.25
Rotation left	38.7	81	0.61
Combined movement A right‡	51.6	55	0.15
Combined movement A left	51.6	81	0.61
Combined movement B right§	64.5	55	0.19
Combined movement B left	53.2	58	0.20
Lateroflexion right	48.4	68	0.38
Lateroflexion left	62.9	52	0.05

*Mean prevalence of limited movements found by examiners A and B.
†Kappa cannot be calculated because the number of limited movements was too small.
‡Combined movement A = extension plus homo-lateral flexion plus homo-lateral rotation
§Combined movement B = lateral flexion combined with a hetero-lateral rotation

Table 3. *Interexaminer agreement of intersegmental mobility of the cervical spine*

	Limited movements† (mean prevalence) %		Agreement (%)		Kappa	
	Right	Left	Right	Left	Right	Left
C0-1	17.7*	17.7*	77*	77*	0.29*	0.29*
C1-2	8.1	11.1	84	90	0.20	0.37
C2-3	30.7	21.0	77	84	0.34	0.63
C3-4	40.3	19.4	85	65	0.20	0.26
C4-5	56.5	22.6	68	68	0.16	−0.09
C5-6	54.9	29.1	61	55	0.17	0.09
C6-7	45.2	20.5	77	48	0.34	0.03
C7-T1	14.6	16.2	74	77	0.08	0.14
T1-2	17.8	21.0	77	84	0.33	0.46

*Tested movement was flexion; both joints tested at the same time; all other tested movements were lateroflexion.
†Mean prevalence of limited movements found by examiners A and B.

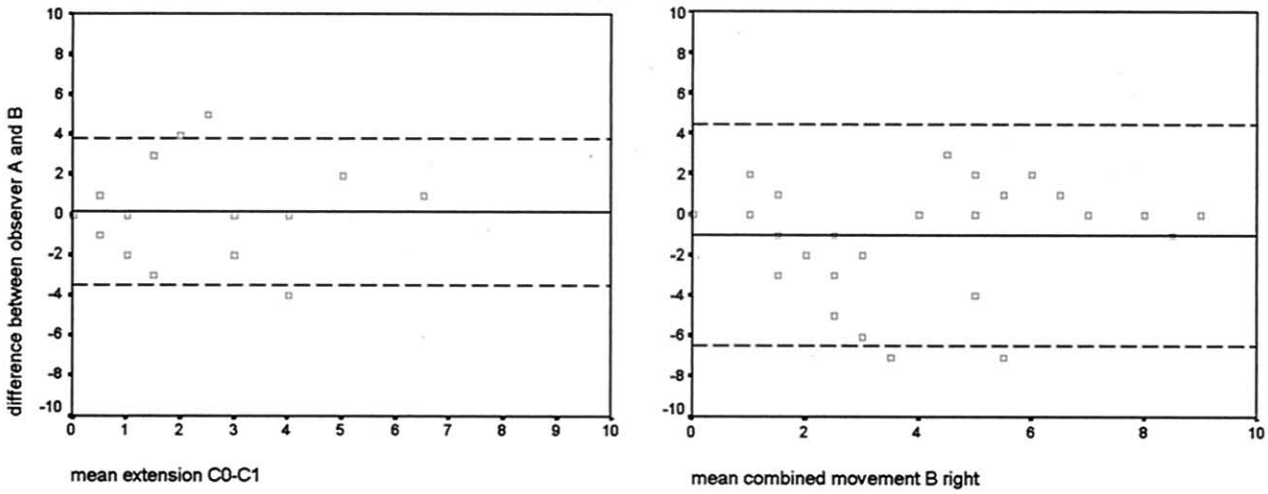


Fig 1. *Agreement of provoked pain scores on a numerical rating scale during assessment of general mobility between examiners. Dotted lines, limits of agreement. Straight lines, mean difference between examiners.*

Table 4. *Interexaminer agreement scores according to the Bland and Altman method and reliability (ICC) scores of provoked pain during assessment of general mobility of the cervical spine*

	Mean difference*	SD of difference	Range of difference	Limits of agreement	Agreement ± 1 point %	Agreement ± 2 points %	ICC
Flexion	0.13	2.22	10	−4.22, 4.48	56.3	75.1	0.63
Extension	0.39	1.96	10	−3.45, 4.23	62.6	81.3	0.71
Flexion C0-1	0.55	2.67	14	−4.68, 5.78	46.9	68.8	0.36
Extension C0-1	0.13	1.86	9	−3.51, 3.77	65.7	75.2	0.56
Rotation right	0.23	2.29	11	−4.26, 4.72	56.3	75.0	0.70
Rotation left	−0.61	2.33	10	−5.18, 3.96	50.0	62.5	0.66
Combined movement A right†	0.06	2.43	9	−4.70, 4.82	50.0	65.6	0.58
Combined movement A left	0.71	2.52	11	−4.23, 5.65	53.0	65.5	0.55
Combined movement B right‡	−1.03	2.78	10	−6.48, 4.42	46.9	69.3	0.54
Combined movement B left	−0.45	2.22	9	−4.80, 3.90	50.0	68.8	0.65
Lateroflexion right	0.58	2.33	12	−3.99, 5.15	56.3	78.2	0.65
Lateroflexion left	−0.19	2.87	12	−5.86, 5.44	50.0	62.5	0.45

ICC, intraclass correlation coefficient.
*Difference in provoked pain score on an 11-point rating scale between examiners A and B.
†Combined movement A = extension plus homolateral flexion plus homolateral rotation.
‡Combined movement B = lateral flexion combined with a hetero-lateral rotation.

74%. The κ ranged from −0.09 to 0.63. Only the levels C2-3 and T1-2 on the left side showed a κ higher than 0.40.

Interexaminer Reproducibility of Provoked Pain Score

The results of interexaminer agreement and reliability for the provoked pain scores during general cervical movements are shown in Table 4. The mean difference between observers A and B was calculated for each movement and varied between −1.03 and 0.58. The limits of agreement were broadest for lateroflexion to the left. The agreement scores within 1 point on the numerical rating scale range from 46.9 % to 65.7%, with a mean agreement of 53.7%. The mean agreement within 2 points is 70.6%. To visualize the results, the best and worst results for the interexaminer agreement are presented in a Bland-Altman plot in Figure 1. All other plots, however, show a distribution between these results. The ICCs for provoked pain scores ranged from 0.36 to 0.71 for the general mobility of the cervical spine, but none of the provoked pain scores reached 0.75.

The agreement of the provoked pain scores during the intersegmental movements are shown in Table 5. The mean difference between observers A and B varied from −1.10 to 0.26, and in most movements, the provoked pain scores for observer B were somewhat higher. The limits of agreement were broadest for C6-7 on the left side. The agreement scores within 1 point of the numerical rating scale range from 40.6% to 75.0%, with a mean agreement of 58%. The mean agreement within 2 points is 76.4%. Again, the best

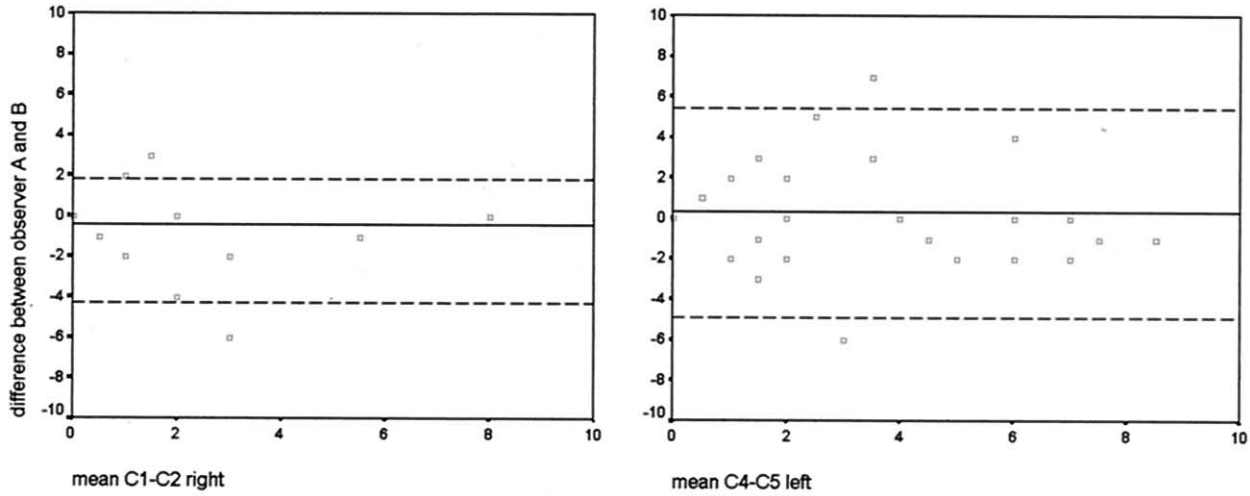


Fig 2. *Agreement of provoked pain scores on a numerical rating scale during assessment of intersegmental mobility between examiners.* Dotted lines, *limits of agreement.* Straight lines, *mean difference between observers.*

Table 5. *Interexaminer agreement scores according to the Bland and Altman method and reliability (ICC) scores of provoked pain during assessment of intersegmental mobility of the cervical spine*

Tested movements	Mean difference*	SD of difference	Range of difference	Limits of agreement	Agreement ± 1 point	Agreement ± 2 points	ICC
C0-1	−0.57	1.45	8	−3.41, 2.27	68.7	87.4	0.73
C1-2 R	−0.48	1.95	9	−4.30, −1.83	75.0	84.4	0.56
C1-2 L	−0.55	2.83	15	−6.10, 5.00	65.6	81.1	0.35
C2-3 R	0.03	2.54	13	−4.94, 5.00	56.4	81.4	0.50
C2-3 L	−0.13	1.87	9	−3.80, 3.53	63.2	82.0	0.78
C3-4 R	−0.52	2.32	11	−5.07, 4.03	62.5	75.1	0.62
C3-4 L	−0.16	2.05	9	−4.18, 3.85	59.4	78.2	0.75
C4-5 R	−0.83	2.19	11	−5.12, 3.46	53.1	78.1	0.62
C4-5 L	0.26	2.64	13	−4.91, 5.43	40.6	68.7	0.55
C5-6 R	−0.87	2.16	10	−5.10, 3.36	62.6	75.1	0.66
C5-6 L	−0.16	2.16	9	−4.30, 4.07	56.3	71.9	0.65
C6-7 R	−0.29	2.53	12	−5.25, 4.67	53.2	72.0	0.59
C6-7 L	−0.35	3.46	16	−7.13, 6.43	47.0	59.5	0.22
C7-T1 R	−1.03	2.59	16	−6.11, 4.05	59.5	75.1	0.45
C7-T1 L	−0.94	2.82	16	−6.47, 4.59	56.3	75.1	0.34
T1-2 R	−1.03	1.43	6	−3.83, 1.77	56.3	81.3	0.80
T1-2 L	−1.10	2.28	10	−5.57, 3.37	56.2	71.8	0.54

ICC, intraclass correlation coefficient; *R,* right; *L,* left.
*Difference in provoked pain score on an 11-point rating scale between examiners A and B.

and the worst results of the interexaminer agreement are plotted in Figure 2.

The ICCs for intersegmental mobility ranged from 0.22 to 0.75. An ICC higher than 0.75 was found for the levels C2-3 and C3-4 on the left side and for T1-2 on the right side.

Discussion

Despite considerable training and the use of a standardized protocol, the results of this study showed that the reproducibility of cervical mobility and pain provoked during mobility assessments was highly variable and overall unacceptable. The assessment of intersegmental mobility showed a slightly better agreement, followed by the agreement for general mobility. However, the κ values were disappointing. The explanation for the differences in agreement and Cohen's κ (the agreement corrected for chance) could be the unequal prevalence of positive or negative findings.[10,17] Fjellner et al,[10] for example, only calculated the κ if the mean prevalence of positive findings for 2 examiners was between 10% and 90%. In this study, the distribution of negative findings (ie, movement not limited) was more than 90% for high cervical flexion and extension.

For the general mobility of the cervical spine, the standard deviation of the difference between provoked pain scores varied between 2 and 3 points. The limits of agreement, which give an indication of the measurement error, varied from 4 to 6 points. These limits of agreement are much too wide to label them as an acceptable agreement. The pain scores for the intersegmental movements show similar limits of agreement as for the general movements, with the exception of a systematic difference between examiner A and examiner B (the scores of examiner B being systematically higher).

An ICC score of 0.75 was defined as an acceptable level of reliability, and in this study, only 3 of the 29 measurements met that criteria.

Studies focusing on the reproducibility of methods to assess the cervical spine are rare. Most studies that have examined the mobility of the cervical spine have reported that it is difficult to achieve a reasonable score for agreement and reliability. Jull et al[21] studied interexaminer agreement in detecting cervical joint dysfunction and reported very high κ values. Viikari-Juntura[13] studied a total assessment of the neck with a conventional neurological examination, palpation of the neck and shoulder region, and clinical tests consisting of 34 items and concluded that only 4 items showed good agreement. Strender et al[12] evaluated the interexaminer agreement of 10 clinical tests and found that only 2 tests had an acceptable agreement. Fjellner et al[10] evaluated a number of clinical tests for the assessment of passive and intersegmental movement and found an acceptable agreement in tests of passive general movement but found an acceptable agreement in a very few of the tests for passive intersegmental movement. Smedmark et al[11] studied 4 tests performed on 61 patients and found a relatively high percentage of agreement but fair to moderate κ values.

The studies carried out by Strender et al[12] and Fjellner et al[10] focused on normal healthy subjects, so the results may therefore not apply to patients with neck pain. The majority of the above mentioned studies reported that better operational definitions and testing procedures were needed.

The assessment with the most reliable score reported in the literature is the foramen compression test ($\kappa = 0.43$),[12] but the combined movement A in the present study, which involves a similar movement, had a low kappa and agreement on the right side. Because different techniques are used in daily practice, it is difficult to make comparisons between studies. Furthermore, only a few studies have focused on patients with neck pain, and the agreement and reliability of provoked pain scores have not been studied. Although the pain score during an assessment is rather subjective, a classification of a dysfunction is made on the basis of the parameters of pain or restricted or limited movement.[2] For most patients, it was very hard to report the same pain score. A higher pain score might have been expected for the second assessment, but no systematic difference was found in the scores. The variation is probably due to the variation between the examiners,[9,22] for example, because of differences in palpation and movements of the same level of the cervical spine and/or to the force used for over-pressure. Earlier studies have suggested to improve reproducibility in daily practice by standardization of the examination protocol. However, even with extensively trained physiotherapists, we found unsatisfying results.

Assessments of general or segmental mobility in daily practice are poorly reproducible; therefore, to diagnose only on the outcome of such an assessment is not recommended. More research is needed in the search for reliable instruments and techniques in daily practice. For research purposes, it is always possible to increase the sample size, which is a strategy to cope with measurements with a large amount of random error.

Conclusion

Despite the use of a standardized protocol to assess general mobility and intersegmental mobility of the cervical spine, it is difficult to achieve reasonable agreement and reliability between 2 examiners. Likewise, the patients are not able to score the same level of provoked pain in 2 assessments within an interval of 15 minutes.

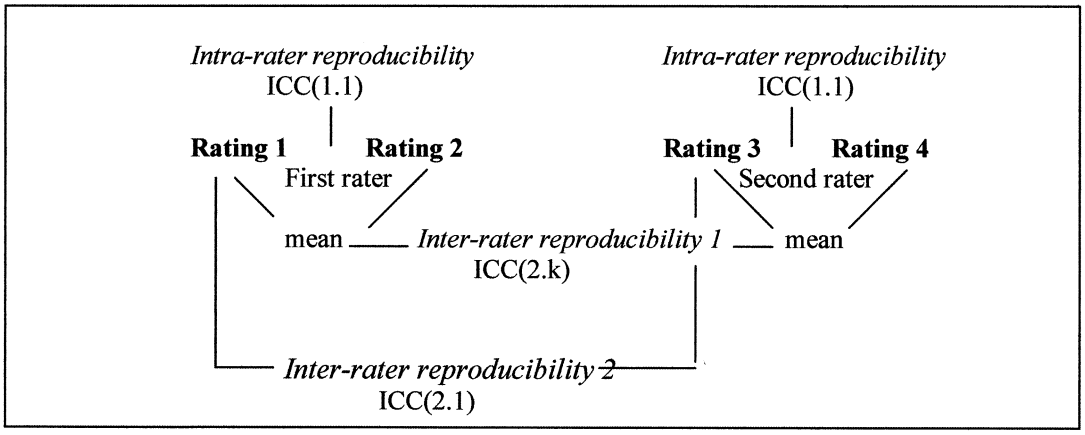
References

1. Borghouts AJ, Janssen HJ, Koes BW, Muris JWM, Metsemakers JFM, Bouter LM. The management of chronic pain in general practice: a retrospective study. Scand J Prim Health Care 1999;17:215-20.
2. Gross AR, Aker PD, Quartly C. Manual therapy in the treatment of neck pain. Rheum Dis Clin North Am 1996;22:579-98.
3. van der El A, Lunacek PBP, Wagemaker AJ. Manual therapy: treatment of the spine (Dutch: Manuele Therapie;wervelkolom behandeling). 2nd ed. Rotterdam, The Netherlands: Manuwel; 1993. p. 329-450.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
5. Brennan P, Silman A. Statistical methods for assessing observer variability in clinical measures. BMJ 1992;304:1491-94.
6. Fleiss JF. The design and analysis of clinical experiments: reliability of measurement. New York: John Wiley and Sons; 1986. p. 1-33.
7. De Vet HCW. Observer reliability and agreement. In: Armitage P, Colton T, editors. Encyclopaedia biostatistica. Vol. 1. Boston: John Wiley; 1998. p. 3123-8.
8. Jull GA. Physiotherapy management of neck pain of mechanical origin. In: Giles LGF, Singer KP, editors. The clinical anatomy and management of back pain series. Oxford: Butterworth-Heinemann; 1997. p. 168-91.
9. Cattrysse E, Swinkels RAHM, Oostendorp RAB, Duquet W. Upper cervical instability: are clinical tests reliable? Man Ther 1997;2:91-7.
10. Fjellner A, Bexander C, Falleij R, Strender LE. Interexaminer reliability in physical examination of the cervical spine. J Manipulative Physiol Ther 1999;22:511-6.

11. Smedmark V, Wallin M, Ardvidsson I. Interexaminer reliability in assessing intervertebral motion of the cervical spine. Man Ther 2000;5:97-100.
12. Strender LE, Lundin M, Nell K. Interexaminer reliability in physical examination of the neck. J Manipulative Physiol Ther 1997;8:516-20.
13. Viikari-Juntura E. Interexaminer reliability of observations in physical examinations of the neck. Phys Ther 1987;10:1526-32.
14. Hoving JL, Koes BW, de Vet HCW, van der Windt DA, Assendelft WJ, van Mameren H, et al. Manual therapy, physical therapy or continued care by the general practitioner for patients with neck pain. A randomized, controlled trial. Ann Intern Med 2002;136:713-22.
15. Vernon H, Mior S. The Neck Disability Index: a study of reliability and validity. J Manipulative Physiol Ther 1991;14: 409-15.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;70:213-20.
17. Streiner D, Norman G. Health measurement scales: a practical guide to their development and use. 2nd ed. Oxford: Oxford University Press; 1995. p. 104-27.
18. Atkinson G, Neville AM. Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Med 1998;26:217-38.
19. de Vet HCW, Beurskens AJHM. Roaming through methodology. VII: reproducibility of measurements. (In Dutch: Dwalingen in de methodologie. VII: reproduceerbaarheid van metingen). Ned Tijdschr Geneeskd 1998;142:2040-42.
20. Kramer MS, Feinstein AR. Clinical biostatistics LIV: the biostatistics of concordance. Clin Pharmacol Ther 1981;29: 11-23b.
21. Jull G, Zito G, Trott P, Potter H, Shirley D, Richardson C. Inter-examiner reliability to detect painful upper cervical joint dysfunction. Aust J Physiother 1997;43:125-9.
22. Gonnella C, Paris SV, Kutner M. Reliability in evaluating passive intervertebral motion. Phys Ther 1982;4:436-44.

APPENDIX

Flow diagram measurements and assessment of intra- and inter-rater reproducibility*



*ICC = Intra-class correlation coefficient, k = number of measurements / raters